DOCUMENT RESUME

ED 302 569 TM 012 572

AUTHOR Cousins, J. Bradley; And Others
TITLE Safeguards against the Misuse of Instruments for the Selection and Appraisal of School Administrators = Mesures pour eviter les abus dans l'utilisation d'instruments de selection et d'evaluation du personnel des administrations scolaires.
PUB DATE Jun 88
NOTE 40p.; Paper presented at the Annual Meeting of the Canadian Association for the Study of Educational Administration (Windsor, Ontario, June 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Administrators; Educational Assessment; Elementary Secondary Education; Evaluation Methods; *Evaluation Utilization; Foreign Countries; Job Performance; *Personnel Evaluation; Principals; Professional Development; Test Construction; *Testing Problems; Test Manuals; *Test Use
IDENTIFIERS Canada; *Selection Tools

ABSTRACT

The legitimate uses of new instruments for the selection and appraisal of principals, factors influencing use of these instruments, and safeguards to minimize misuse are discussed. Performance appraisal systems in education have been used legitimately to ensure that principals meet minimum performance standards, and for promotion. Another valuable function is in meeting professional development needs of principals. Data from appraisal instruments may be incorrectly or inadequately used. Use and misuse of instruments are influenced by: (1) the conceptual framework used; (2) the perspective of the test developer; (3) the perspective of the practitioner; and (4) links to factors influencing use. Safeguards that instrument developers and practitioners alike may use include: (1) an explicit and clear technical manual; (2) a standard system for processing data; (3) extensive field training of practitioners; (4) use of instruments as part of a battery of procedures for evaluation; and (5) caution about claims that any one instrument can define the ideal principal. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED302569

Safeguards Against the Misuse of Instruments for the Selection and Appraisal of School Administrators

Mesures pour éviter les abus dans l'utilisation d'instruments de sélection et d'évaluation du personnel des adminstrations scolaires

J. Bradley Cousins, Paul T. Begley and Kenneth A. Leithwood

Centre for Principal Development
The Ontario Institute for Studies in Education

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DR. J. BRADLEY COUSINS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the
Canadian Association for the Study of Educational Administration
Windsor, June, 1988

TM012572

BEST COPY AVAILABLE

Safeguards Against the Misuse of Instruments for the Selection and Appraisal of School Administrators

Mesures pour éviter les abus dans l'utilisation d'instruments de sélection et d'évaluation du personnel des administrations scolaires

by
J.B. Cousins, P.T. Begley & K.A. Leithwood
Centre for Principal Development
The Ontario Institute for Studies in Education

Public demands for due process and accountability and rapidly increasing retirement patterns in education have heightened school systems' needs for personnel selection and appraisal procedures that are fair, effective and efficient (i.e., cost effective). Because the role of school administrators has been shown to account for a substantial portion of the variation in school effectiveness, the implications of these trends are crucial for school systems that seriously wish to implement planned educational change.

In response to these trends, several new instruments have been developed for the purposes of use for selection and appraisal of principals. The growing availability of such measurement devices carries with it the possibility that they will be in some way misused. The objective of this paper is to examine legitimate uses of such instruments, probable types of non-use, misuse, and abuse, and ways in which misuses can be minimized, if not avoided. These issues will be examined from the perspectives of both instrument developers and educational practitioners. To be more specific, the objectives are to identify: (1) ways in which instruments might legitimately be used in the context of school administrator selection and appraisal; (2) factors that are likely to influence the use of instruments for such purposes; and (3) safeguards that instrument developers and educational practitioners might employ in order to ensure that misuses of such instruments are minimized.

Context

Current demographic and socio-educational trends have led to rather dramatic implications for school administration. From the demographic perspective, available

data suggest that a significant turnover in school leaders may soon be expected due to an aging faculty, early retirement incentives, and the like. Leithwood and Begley (1985) cite Ontario and Alberta data that predict as many as 40 percent of current principals will be eligible for retirement (age 55) by 1995. A more alarming statistic reported by Peterson, Marshall and Grier (1987) indicated that "more than 60 percent of administrators [will be] retiring by the end of the decade" (p. 47). Regardless of the precision of estimates, the importance of training, selecting and appointing school administrators will increase significantly in the very near future.

As identified by Leithwood and Begley, among others, at least one major implication of these trends concerns the adequacy of current selection practices. They found that very few systems have formal, printed policies regarding school administrator selection and that in general seven criteria are commonly used: (1) academic training; (2) teaching experience; (3) good health; (4) administrative experience; (5) recommendations from colleagues; (6) inservice training; and (7) personal attributes. The adequacy of these criteria is suspect given that they are not derived from a specific image of the effective administrator. Moreover, even if they were, the changing nature of the principals' role would potentially make them obsolete.

Much of the empirical literature associated with the effective schools movement has indicated an increasingly important role for the principal as an instructional leader. As mentioned earlier, the principal has been shown to explain a significant proportion of the variation in school effectiveness (Robinson & Block, 1982; Hallinger & Murphy, 1987; Leithwood & Montgomery, 1982; 1986). One implication of this finding was stated aptly by Leithwood and Begley.

> Given the large number of principals in Canada who will retire in the next five to ten years, selecting the most effective candidates to replace them may be viewed as a major strategy for school improvement (1985, p. 15).

Essentially the process of principal selection is a school system administrative function that may or may not involve the systematic collection of information. This is

also the case for some aspects of performance appraisal. Due to a loss by schools of public trust, increased demands for accountability have been placed on the educational system. As a consequence, performance appraisal systems in education have been used for such administrative purposes as ensuring principals meet minimum performance standards, and promotion.

Another valuable function of performance appraisal systems concerns meeting the professional development needs of principals. The use of data concerning their own performance for such purposes has been acknowledged by principals and other educational administrators as being very important. Yet data show that actual use for this purpose has been extremely limited. Even though principals appear to take appraisal processes seriously, the impact of such data on their performance has been negligible (Cousins, in preparation; Duke & Stiggins, 1985; Lawton, Hickox, Leithwood & Musella, 1986). These data are similar to those of Leithwood and Montgomery (1982) who found that the majority of principals believe that instructional leadership is important, yet fewer than half do anything about it. Use of time, lack of knowledge and skill, and absence of planning and reflection were suggested as reasons underlying these observations.

Given current retirement patterns and principals' acknowledged view of appraisal data as being important for professional growth, opportunities abound for school systems to effect planned educational change through school administrator selection, promotion and appointment, and professional development through appraisal. In keeping with current demands for rationalized selection and appraisal procedures several new measurement devices have been developed to serve both administrative decision-making and developmental (i.e., performance improvement) needs.

Some instruments use performance assessment techniques (primarily in "assessment centre" settings such as those affiliated with the National Association of Secondary School Principals -- NASSP) to assess principal (or principal candidate)

performance on key, simulated, job-related tasks. Stiggins (1987) provided useful advice for constructing performance assessment instruments. Other instruments have been designed to evaluate the effectiveness of assessment centers in achieving their objectives (e.g., Schmitt, Noe, Meritt, Fitzgerald & Jorgensen, 1983). These instruments are typically behaviour observation rating scales that can be completed by the principal (or candidate) in a self-evaluation context or by colleagues (supervisors, peers, subordinates) with some knowledge of the individual's professional behaviours. While some such devices have been adapted to education from other settings (e.g., business and industry, Pitner & Hocevar, 1987) to measure principal effectiveness, others have been developed entirely within the context of school leadership (e.g., Cousins & Leithwood, 1987; Hallinger & Murphy, 1984; 1987; Leithwood & Montgomery, 1986). More recently, Hall (1988) and Vandenberghe (1988) have collaborated on the development of an instrument to measure principals' change facilitator style.

Concepts of Use

The use of instrumentation for administrative decision-making and professional development can be conceptualized within the bounds of evaluation utilization or, more generally, knowledge utilization. Conventional conceptions of knowledge utilization (e.g., Alkin, Daillak & White, 1979; Weiss, 1981) have relied on a continuum that ranges from instrumental uses (support for decision-makers' decisions) to uses for conceptual development (the contribution of knowledge to user learning or educational outcomes). As both Alkin et al. and Weiss observed, much of the initial research on use employed the instrumental conception and this severely restricted the range of observed uses of knowledge. Patton and his associates (Patton et al., 1977) were among the first to recognize the inadequacy of conceptualizations of use in strictly instrumental terms. This recognition has stimulated research on use considerably as indicated by the proliferation of empirical research (e.g., Cousins & Leithwood, 1986, reviewed 65 empirical studies from the period 1970-1985. Many of these studies were reported in the last five years of that period).

More recently, an even broader conception of the utilization process has been suggested (e.g., Cousins & Leithwood, 1986; Huberman, 1987; Kennedy, 1983; 1984) -- one that defines use as information processing by the user. According to this conception the mere psychological processing of information (e.g., evaluation data, test scores, other forms of evidence, etc.) without necessarily informing decisions, dictating actions to be taken or changing thought processes, constitutes use. For example, a user could read and fully comprehend the contents of an evaluation report but not learn anything from it nor base decisions on it. The framework developed by Cousins and Leithwood (1986) incorporated all three types of use in its definition of the dependent variable, utilization (see Figure 1). According to this conceptualization, evaluation data could be used to support discrete decisions (e.g., staffing, program management, program funding) or to educate decision-makers about aspects of the object of evaluation (e.g., nature of program impact, components of programs explaining outcomes, etc.). Prior to either type of use, however, data must first be cognitively processed by decision-makers (e.g., given serious consideration) which might also result in it being discarded from further attention.

Among others, Duke and Stiggins (1985) distinguished between two purposes of evaluation concerning principals: accountability (summative evaluation) and professional development (formative evaluation). Meeting the first type of purpose might be construed as "utilization as decision" because the evaluation system is designed to support the administrative decision process. In particular, such systems can provide knowledge for decisions about selecting from a pool of principal candidates, individuals suitable for the role, promoting such individuals into positions of added responsibility, and ensuring individuals are functioning in their roles at acceptable levels of competence. Thus, data derived from instruments might be used legitimately to support administrative decision-making. However, in a recent Ontario survey conducted by Musella and Lawton (1986) no such evidence was obtained. Information collected for the purposes of selection and promotion typically included letter of application, application

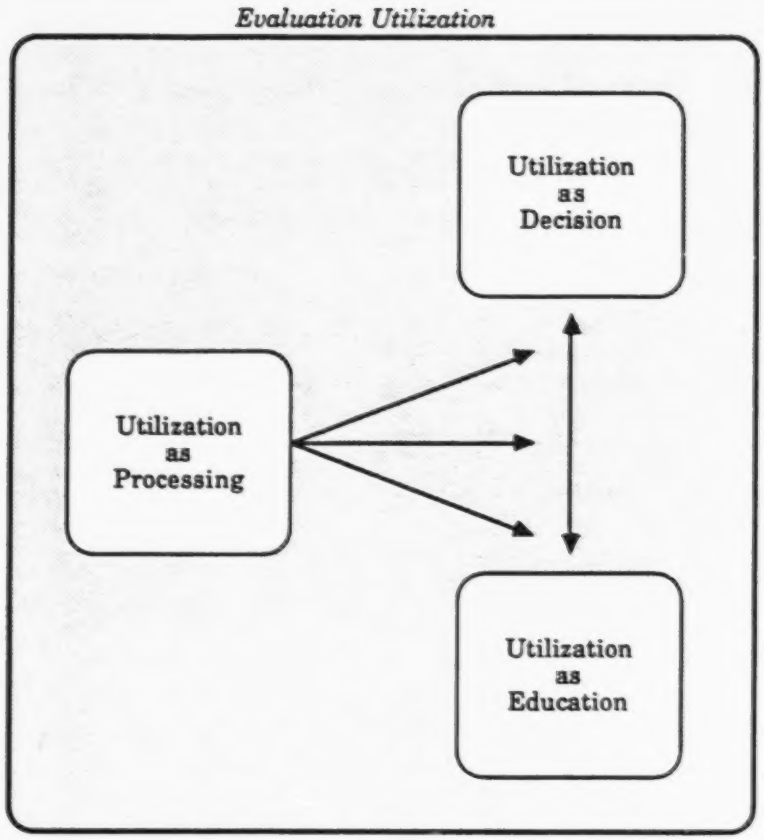

Figure 1

Conception of Utilization

forms, written reports from applicants and supervisors, references, and statements of educational philosophy. Until recently the availability of instruments for such purposes has been very limited but with the recent development of new instruments and the growing need for quick information it seems likely that such data will soon be incorporated into selection and appraisal practices. Indeed some systems are currently using behaviour rating scales to identify potential principal candidates who subsequently undergo more thorough selection procedures (e.g., interviews).

The other main purpose of principal evaluation professional concerns conceptual development or education. In this context the ultimate users of the data are the principals themselves and use is defined in terms of conceptual development manifest in performance improvement. Recent evidence has shown that although this purpose for principal evaluation is thought to be highly important, in actuality, there is a very weak link between evaluative information and professional development (Duke & Stiggins, 1985). Also, Lawton, et al. (1986) found that even though principals took appraisal processes quite seriously, the impact on the improvement of performance was limited, if not negligible. Similarly, Cousins (forthcoming) found extremely little "new learning" was attributable to appraisals; reinforcement of current knowledge was the more common type of conceptual development but only modest evidence of such development was available. With continued emphasis on public accountability and school systems' desires for excellence in education it seems likely that interest in instruments for purposes of professional development will continue for the forseeable future. Some instruments (e.g., The Principal Profile, Leithwood & Montgomery, 1986) were designed with this specific purpose in mind. One appraisal model that appears to be gaining some popularity (we are aware of at least three school systems) involves the explicit use of a behavioural rating instrument by the principal, his supervisor and subordinates and the subsequent comparison and discussion of results in a post appraisal conference. This approach is more concerned with formative benefits rather than summative judgments about the principal.

Misuse of Data

In the previous paragraphs we have described two main purposes of principal evaluation and legitimate uses of measuring devices for such purposes. Given these intentions, what are the possible or probable misuses of such devices? Alkin and Coyle (1988) have recently provided some very interesting ideas regarding the misutilization of evaluation that are helpful for framing the present discussion.

Patton (1988, cited by Alkin and Coyle) identified two separate dimensions, utilization and misutilization, that are relevant. Utilization might be thought of as a continuum with use for decision-making or conceptual development (depending on purpose) at one extreme and non-utilization (i.e., failure for users to process data) at the other. Similarly, variation in misuse may be plotted along a continuum ranging from misutilization to justifiable action.[1] As an heuristic for differentiating types of misuse, it might be helpful to consider these dimensions to be completely independent of one another as shown in Figure 2.

Alkin and Coyle, in their attempt to define misutilization of evaluation data, describe several variations that may be distinguished. Data derived from an instrument may be expected to vary in terms of their technical quality. It may be more or less free of measurement error (reliable), and may have been derived from multiple raters such as principals, superiors, teachers (valid). If information about an individual is known to be of superior technical quality but is surpressed by a particular potential user (e.g., selection committee member) for whatever reason we have an instance of what Alkin and Coyle refer to as <u>abuse</u>: a clear case of non-utilization of the data that can readily be described as misutilization.

But some instances of non-utilization may be legitimate. For example, if an administrator is aware that the results of an assessment are technically flawed or

---

[1]Patton uses the term 'non-misutilization' to indicate justifiable action. We have substituted purely in the interest of avoiding the double negative.

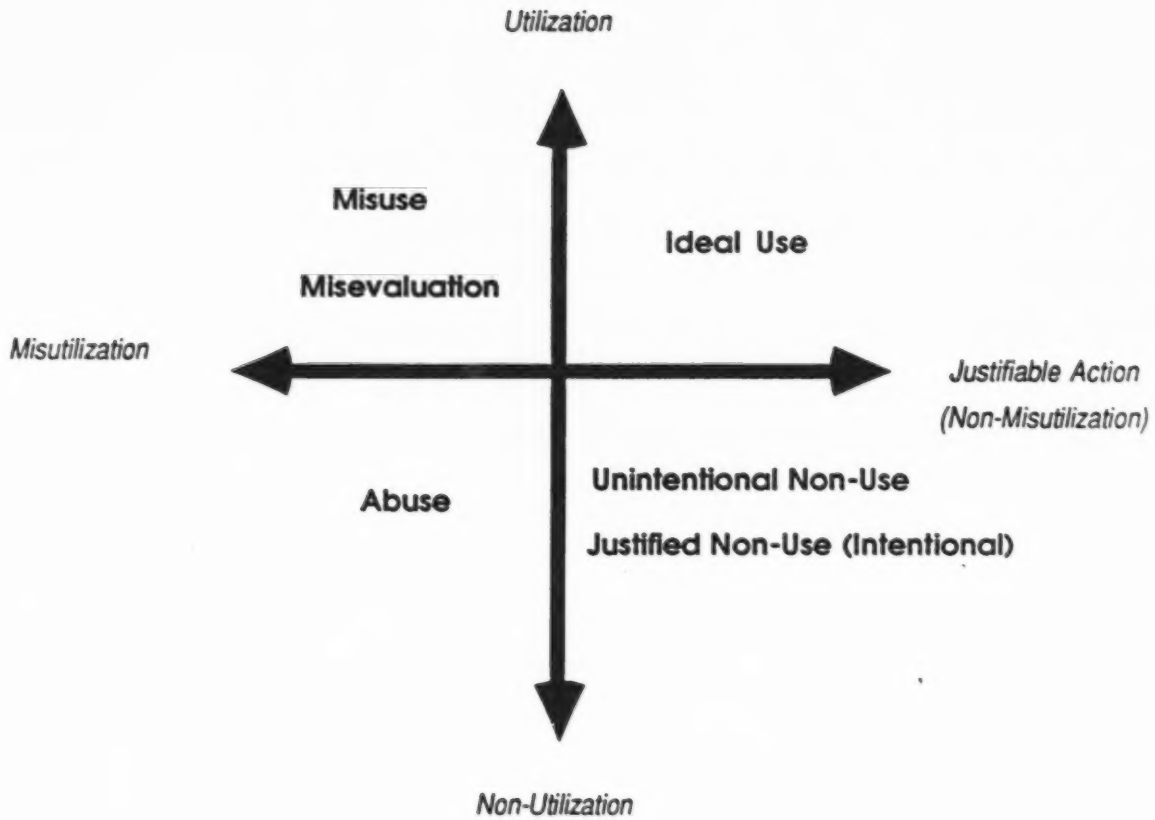

Figure 2

Conception of Misutilization

erroneous, he or she would be justified in not incorporating the information into the decision-making process. This Alkin and Coyle referred to as justified non-use and may be considered an appropriate and responsible action. If, on the other hand, the data are of sufficient technical quality but potential users are unaware of its existence, or inadvertently fail to process the information, we have a case of unintentional non-use. This particular example cannot be considered misutilization although it is certainly not a desirable outcome for systematic decision-makers.

The best of possible outcomes is that instrument data of sufficiently adequate quality are processed by users and subsequently used for either conceptual development and/or decision-making purposes (depending on intentions). This, there is strong agreement, would be a condition of ideal use. A less satisfactory, but nonetheless legitimate form of use, would be for potential users to cognitively process the instrument results but, subsequently, not to learn from them nor base decisions on them. There may, in fact, be legitimate reasons (e.g., competing information that is more compelling) for only limited use. For example, a principal might fully understand the results of an instrument used for appraisal but due to prior knowledge not learn anything regarding performance improvement.

Finally, Alkin and Coyle differentiate between two types of misutilization when data are processed. The first they described as misuse, a term that corresponds to the deliberate manipulation of, say, instrument scores to serve some particular purpose (e.g., support or non-support for a principal candidate). Clearly, this situation is an example of intentional misutilization of the data; the data are used but in an inappropriate fashion. The second type of misutilization was termed misevaluation and its sources touch on the issue of who has responsibility for misuse -- the potential user or the instrument developer? In this case, a possible outcome would be that test developers had not taken the necessary steps to prevent misutilization. Incomplete scoring information, absence of normative data, poor administrative instructions, and the like, are possible sources of error that could ultimately lead to misevaluation. Of course, the

responsibility for misevaluation need not necessarily reside with the instrument developer but could be attributed to individuals that administer the instruments through such actions as careless administration, scoring, and so forth.

We have examined a number of possible outcomes concerning uses and misuses of instrument data in the context of school administrator selection and appraisal procedures. The discussion now turns to an examination of factors that potentially bear upon, or otherwise influence these outcomes. Relevant issues will be examined from the perspective of both instrument developers and practitioners but first, as an organizing structure, the Cousins-Leithwood (1986) evaluation utilization framework is described. Ultimately, a series of safeguards against potential misuses of instrument data are suggested.

Factors Influencing Use and Misuse of Instruments

(a) Framework

The conceptual framework used (see Figure 3) has been described in detail elsewhere (Cousins & Leithwood, 1986; Cousins, forthcoming) and is grounded in a considerable body of empirical research (65 studies) about the utilization of evaluation data. Factors in the framework are defined as the circumstances or conditions that influence the extent to which evaluation data are used. Cousins and Leithwood found that factors could be categorized according to one of two major hypothetical dimensions: characteristics of the *evaluation implementation* and characteristics of the *decision or policy setting*. These dimensions are suggested to be correlated and to interact with one another to produce effects (inhibiting, stimulating) on use. Six factors are encompassed by each of the two hypothetical dimensions. Those associated with the evaluation implementation are:

- *Evaluation quality*. Characteristics of the evaluation process including the sophistication of methods, rigor, and type of evaluation model. An evaluation that attempts to link program components to program outcomes, for example, is considered to be more sophisticated than one that merely describes outcomes.

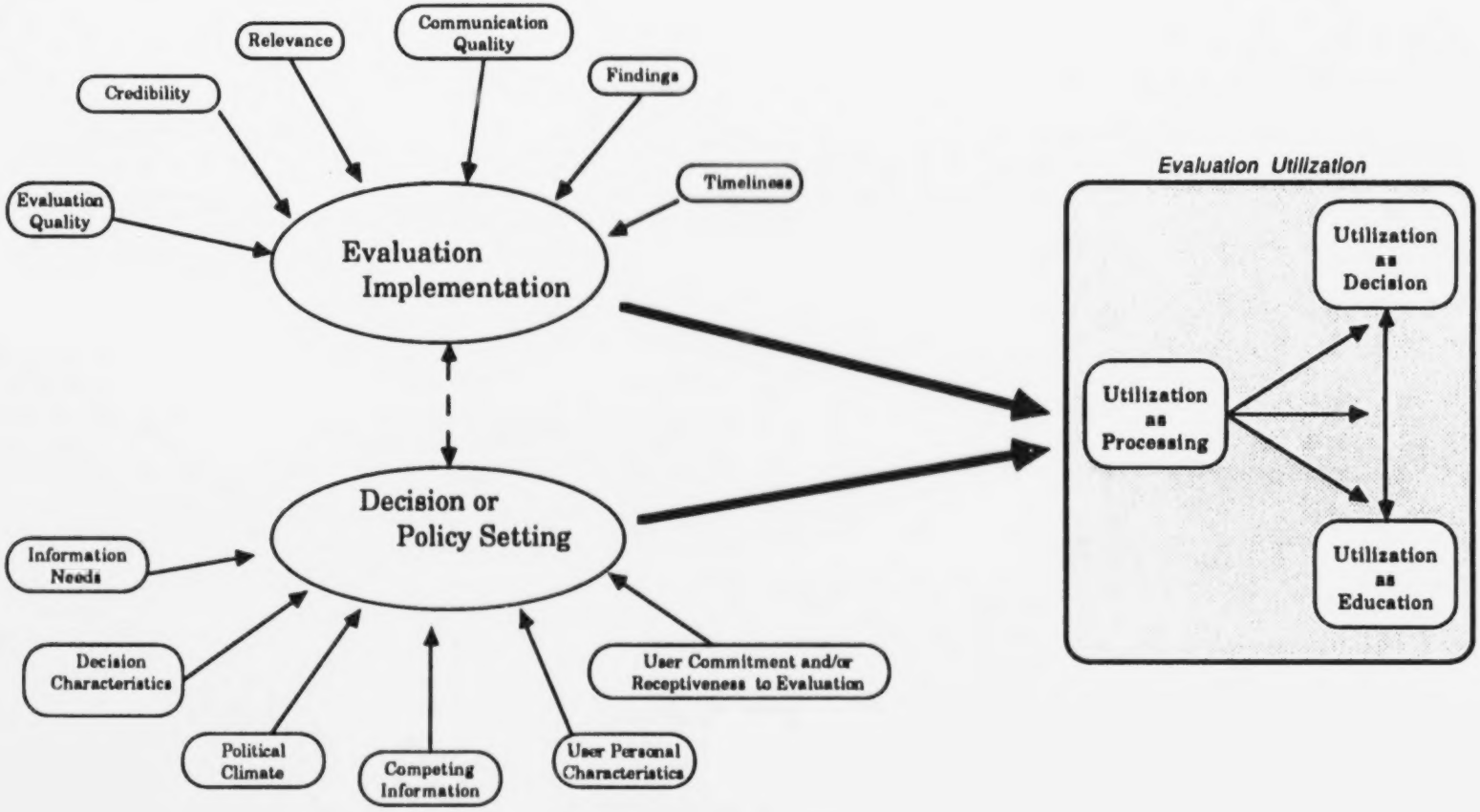

Figure 3

Evaluation Utilization Conceptual Framework

- *Credibility*. Credibility of the evaluator and/or the evaluation process defined by believability, objectivity, appropriateness of evaluative criteria, and the like. A well-seasoned evaluator with a proven track record is attributed higher levels of credibility than a novice, for example.

- *Relevance*. The relevance of the evaluation to the information needs of decision-makers according to: (1) the purposes of the evaluation and (2) the organizational location of the evaluator. Do the purposes of the evaluation meet the explicit and implicit needs of the audience(s) for whom the evaluation is conducted? Do evaluators working within the organization tend to produce evaluations that are more relevant?

- *Communication quality*. Quality and/or clarity of the dissemination of results to the evaluation audience(s) according to characteristics such as the style of the report and the propensity of the evaluator to advocate its results. For example, is the report presented orally and/or in written form and does the evaluator follow-up the presentation with clarification?

- *Findings*. The nature of the evaluation findings was defined by their positive or negative valence (e.g., judgments about whether the program is meeting its objectives), their consistency with the expectations of the evaluation audience(s), their value for decision-making, and the like. To what extent are findings predictable to decision-makers?

- *Timeliness*. The point in time at which evaluation results are disseminated to decision-makers relative to impending decision(s). Are the results presented too late to have an impact on the decision process?

Factors associated with the decision or policy setting are:

- *Information needs*. The type of information sought, number of evaluation audience(s) with differing information needs, time pressure and perceived need for evaluation. To what extent are explicit and implicit needs for evaluation information shared among different audiences?

- *Decision characteristics*: Characteristics of decisions associated with the evaluation problem including decision impact area, type of decision, program novelty and significance of the decision, among other examples. Decisions regarding politically sensitive or controversial issues are of relatively high significance.

- *Political climate*. Characteristics associated with political climate such as political orientation of commissioners of the evaluation, dependence of decision-makers on external sponsors, inter- and intra-organizational rivalries, budget fights and power struggles. Is it politically prudent for decision-makers to decide in a manner that is consistent with the evaluation

results?

- *Competing information*. Information from sources beyond the evaluation relevant to the research problem and competing with evaluation data to inform decisions. Personal experience, informal observations made by decision-makers and working knowledge are examples.

- *User personal characteristics*. Decision-makers' organizational role, information-processing style, organizational experience, and social characteristics, among other variables. Decision-makers who carefully plan for the future and take preventative actions are distinguished from "crisis managers" who operate on more of a "reactive" basis.

- *User commitment and/or receptiveness to evaluation*. Extent to which decision-makers are open-minded about decisions and the evaluation findings. Are the decision-makers dogmatic about the decision? Are they predisposed to attitudes about the utility of evaluation?

Although these factors were found to impact on the extent to which evaluation data were used it seems likely that they are relevant also to questions of the misuse of evaluation, or in this case, instrument data.

(b) The Instrument Developer's Perspective

From the perspective of instrument developers various issues emerge concerning the use of instruments for selection and appraisal purposes. These issues are of two general but interconnected types: (1) technical adequacy of the instrument and (2) determination of responsibility for valid use. What are the requirements of such an instrument? What properties are desirable for it to have if valid inferences about the performance of school administrators are to be made? In response to these questions several issue may be categorized under the technical adequacy category.

Issue 1: Construct Validity

A particularly important aspect of validity is the concept of *construct* validity: does the instrument measure what it is intended to measure? For example, what is the correspondence of the observed behaviours as reflected by scores on the instrument with the underlying performance construct of principal effectiveness? Clearly, if the

instrument is to be useful for making valid inferences about performance and subsequent judgments about plans for improvement or administrative decision-making it must be shown to measure what it purports to measure.

Several researchers (e.g., Leithwood & Montgomery, 1986; Pitner & Hocevar, 1987) have shown that the construct of principal effectiveness is multidimensional. For that reason, instruments that reduce an individual's performance to a single score are limited representations of the construct. A more appropriate representation would be a profile or scores on multiple dimensions.

Issue 2: Predictive Validity

A second important validity issue has to do with the utility of the instrument for administrative decision-making about, for example, selection or promotion. The instrument must be shown to be able to predict performance as reflected by some criterion measure at some future point. This type of validity is commonly known as a special case of *criterion-related* validity, namely *predictive* validity. The greater the predictive ability of an instrument, the more valid it would be for making score-based inferences about selection and promotion to positions of higher responsibility. The instrument must also be able to predict scores on a criterion measure administered at roughly the same time. Such data would reflect on another case of criterion-related validity called *concurrent* validity. The problem becomes one of identifying a suitable criterion. Several instruments are available (e.g., Hallinger & Murphy, 1984; Pitner & Hocevar, 1987; Schmitt et al., 1983) but the extent to which they have been shown to be valid measures of effective principal practice is unclear. Other possibilities for criterion measures include ratings based on behavioural observations of supervisors, peers or subordinates but such ratings would likely have limited reliability. Still other types of measures might include actual or simulated performance assessment scores based on modules or packages such as that developed by Stiggins (1987). Again, the validity of these measures is in question and practical concerns become apparent.

Issue 3: Content Validity

The content coverage of an instrument must be shown to be a representative sample of the domain of performance in question: for example, effective school leadership. One image of effectiveness in the role is consistent with a planned change framework where a premium is placed on the establishment of a clear image of goals and objectives. If shown to cover what is meant by effectiveness in the role of principal, an instrument when used for appraisal purposes would have the potential to assess the extent to which a gap exists between observed or actual performance and ideal performance. Moreover, it could be useful for identifying potential obstacles to growth by principals in performance toward effective practice. Validity considered in these terms is generally called *content* validity.

Some authors are of the opinion that construct validity is the only legitimate form of validity to be considered. In his comprehensive review of the topic Cronbach (1971) stated clearly that it is not a test (or instrument in the present case) that becomes validated but the inferences made based upon test scores. For that reason, theorists such as Messick (1981) suggest that "so called content validity is not validity at all" (p. 11) and that content coverage has more to do with test construction than with validating inferences based upon test scores. He asserted that it may be more harmful than helpful to speak of different types of validity. Cronbach (1980) similarly stated that test validation requires many different types of evidence and "all validation is one, and in a sense all is construct validation" (p. 99). Yalow and Popham (1983), on the other hand, argued that although issues such as content coverage are important to these critics, quibbling over terminology may diminish the perceived importance of content coverage in the long run. They asserted that

> If a test is constructed so that it constitutes a representative sample of the domain of interest, then we expect the examinees score on the test reflects how the examinee will perform in the domain of interest .... Appropriate content coverage is the cornerstone of defensible test construction. (p. 11)

We concur fully with Yalow and Popham that defensible interpretations of test

scores cannot occur without the explicit demonstration of appropriate content coverage. Content validity in our view is a legitimate form of validity that needs to be addressed in the early stages of an instrument's developmental process.

Issue 4: Normative Data

An instrument may be shown to be valid in several respects (as outlined above) and yet be limited by the sample from which validation data were collected. It is important to demonstrate that conclusions about validity are generalizable beyond a restrictive sample of principals or principal candidates prior to making the assumption that score-based inferences are widely tenable. Inspection for, and publication of, differences in norms due to regional, cultural, and other demographic variables will help users to make decisions about local, regional and national performance standard setting.

Issue 5: Reliability

A clear requirement of an instrument is that it be shown to be reliable. With a large sample, the assessment of an instrument's internal validity is possible. Such data may even suggest further refinement of the instrument according to item-total and item-subtest (dimension) correlation indicators. Split-half reliability techniques would also provide useful information about internal consistency. Another important reliability issue concerns the extent to which scores are stable. A test-retest paradigm should provide an appropriate test. We would expect scores on the instrument to correlate with scores on the instrument obtained from the same individuals after a prescribed period of time had elapsed.

Issue 6: Errors in Estimation

Intimately related to the topic of reliability is a discussion of various types of measurement error. Landy and Farr (1983) summarize a variety of measurement errors that should be assessed. These include errors of:

- *leniency* or the extent to which raters are unduly lenient or severe with their ratings of an individual;

- *central tendency* or the extent to which individuals are unwilling to use extremes on rating scales;

- *halo* or the tendency for raters to form a global impression of an individual and to rate him or her accordingly thereby not differentiating over items;

- and *inter-rater reliability* or the extent to which ratings of one individual by more than one rater correlate.

These and other important technical issues (see Bernardin & Beatty, 1984 for a comprehensive summary) must be addressed if valid and reliable score-based inferences are to be made using an instrument.

Issue 7: Responsibility for Use

Establishing the reliability and validity of an instrument will likely influence favourably the extent to which practitioners will use it to support discrete decisions or to foster conceptual growth. However, such evidence will do nothing to ensure that valid uses are made of the instrument.

Critics such as Messick (1981) make it quite clear that historically the responsibility for the valid use of a test lies in the hands of the person that interprets it. He states that "a heavy ethical burden thereby falls on the user" (p. 19). Cronbach (1980) had similar sentiments. "Though the developer of a test should help the user in any practical way, validation is the interpreter's responsibility" (p. 99). Indeed, the the Standards for Educational and Psychological Test Use (AERA et al., 1985) reiterate the same message.

What then is the responsibility of the instrument developer? Or perhaps more to the point, what can such individuals do to foster valid use of an instrument? Instrument developers cannot be absolved of all responsibility for an instrument once it has been produced and shown to be reliable and valid. Steps should be taken to prevent misuse in situations where such outcomes are either predictable or have been observed.

(c) The Practitioner's Perspective

A number of issues become evident when one considers the use of instruments for administrator selection and appraisal from the perspective of the user. In this section of the paper several of these issues are examined from the educational practitioner's perspective.

Before beginning the discussion it is useful to be clear about the use of some terms. Note that for the purposes of this paper, the term "user" can refer to individuals or groups that administer selection and appraisal instruments as well those people being evaluated or selected. Whenever appropriate, the variations in these two points of view are compared and contrasted.

The problems or issues which become evident from the users' perspective of instruments for selection and appraisal are of five types and can be grouped in two distinct clusters. Four of the problems relate to the nature of the environment where such instruments are likely to be used and the final problem type has to do with the perceptions and typical orientations of the users of such instruments.

The problems associated with the first cluster of issues relating to environmental concerns include the following:

- the dynamic and continually evolving role responsibilities of the school administrator;
- the essentially regressive, dependence fostering and often short-lived relevance of pre-determined appraisal and selection criteria;
- the widely divergent contextual or situational characteristics which can be encountered in even one community of the same school district;
- the time-constrained, reactive and non-reflective nature of the typical instrument users' workplace.

The final problem type has to do with perceptions and basic orientations of the user described by the preferred administrative style or work orientation.

Issue 8: Evolving Role of School Administrators

This paper began with a discussion of some social trends which have had an impact on education generally and have intensified the need for rationalized personnel selection procedures and systematic performance appraisal procedures. These trends have also had an impact on the actions, responsibilities and expectations associated with school administrators. In particular, the role of the school principal can be seen as having evolved steadily over the last two decades in response to various educational trends and social pressures. Presumably that role will continue to evolve much as it has in the past in response to pressures which cannot always be reliably predicted. This raises the following question. Can we expect selection and appraisal instruments based on past or present practice, however exemplary, to identify appropriate school administrators for the future?

Leithwood and Montgomery identified four distinct stages of growth in The Principal Profile (1986). These four levels of performance provide a useful historical perspective on the evolution of the role. The levels are labeled Administrator, Humanitarian, Program Manager, and Systematic Problem-solver. Each stage describes a relatively complex image of the role along four dimensions of behaviour (Goals, Factors, Strategies, Decision-Making) and illustrates fairly well the evolution of the principal's role since the 1960's. For example, the image of the effective principal during the 1960's was essentially that of the building manager concerned with maintaining a smoothly operating organization and keeping up appearances; hence the Administrator. The increased concern for individual expression and good interpersonal relations characteristic of the early 1970's is manifested in the Humanitarian. Humanitarians augment the traditional building manager's role with an over-riding concern for the quality of staff, student and, to a modest extent, community relations. As mentioned previously, in the late 1970's the following social trends became evident: an erosion of public trust, a demand for more accountability, the development of multiple interest groups within the greater school community and a much more

diversified curriculum. These trends produced a pressure for more centralized control of education and the steady generation of new policies and curriculum documents requiring implementation by school administrators. These pressures created the need for Program Managers, an orientation which still characterizes many effective principals working in schools today.

The fourth and final (at least for the time being) stage of growth in the Profile, the Systematic Problem-solver, represents our best notions as to what is an appropriate image for the principal's role under present and some emerging circumstances experienced by school administrators. This image of the role constitutes a subtle variation on the Program Manager that was so well suited to the late 1970's and early 1980's. Systematic Problem-Solvers are characterized as having a primary concern for the needs of students rather than building management, the quality of interpersonal relationships, or program and policy implementation. Such a focus allows them to deal in a rational way with the overload of implementation tasks currently reported by many school administrators. Their actions are focussed by the individual student needs they perceive, not just the sometimes strident demands of competing special interest groups. Furthermore, these rare individuals are energetic, highly skilled, and self-motivated entrepreneurs who can often identify and mobilize people and resources despite an increasingly resource constrained environment. Unfortunately, according to Trider and Leithwood (1988), very few principals manifest such an orientation at this time.

We know that a few principals with systematic problem-solving orientations do exist and that they probably represent the state of the art in the principalship. We cannot be sure, however, whether it is realistic to expect many people to attain that level of expertise. Systematic Problem-solvers are highly productive, entrepreneurial and extremely energetic workers not frequently encountered in any line of employment. More to the point, because the role of the principal continues to evolve all the time in response to new social trends and issues, can we even be sure what characteristics the effective principals of the future will require? This produces an implication for users of

selection and appraisal procedures. Instruments based on exemplary practice probably describe, at best, ideal current standards of practice which may or may not be attainable by significant numbers of practitioners, nor even appropriate to the needs of the education system such as they may emerge to be in the future.

Issue 9: Using Pre-Determined Selection/Appraisal Criteria

In a recently published thoughtpiece, What's Worth Fighting For In The Principalship, Michael Fullan (1988) develops several notions relating to the role of the school administrator. Several have implications for the design and use of instruments for the selection and appraisal of administrators. He describes the "nonrational" and chaotic world of the school administrator which must be recognized as a pervasive circumstance of their work environment. He also develops the notion of dependency; something he believes should be avoided by school administrators.

Dependency is defined in Fullan's monograph as "one's actions (being) predominantly shaped, however unintentionally, by events and/or actions or directions of others" (1988, p. vi). He then presents a compelling argument for the empowerment of school administrators by reducing their dependency on pre-determined criteria for guiding or assessing the merit of their professional actions. With specific references to the highest levels of effectiveness espoused by such frameworks as The Principal Profile, Fullan questions whether "a well worked out profile in the hands of superordinates, who themselves may not be systematic problem-solvers, create(s) a sense of dependency among principals as they attempt to measure up." (1988, p. 9)

Fullan's concerns are similar in many respects to those identified in the previous section of this paper. Using an image of effective practice based on criteria derived from the past, or even present exemplary practice, for selection or appraisal of school administrators is essentially limiting and counterproductive. As an alternative, Fullan advocates that new principals adopt or aspire to a set of essential concepts or skills and concepts. These he suggests might include integrity, initiative, internal locus of control,

risk taking, perpetual learning, and political astuteness. However, it seems unlikely that such constructs can be easily measured. For the time being anyway, the only promising indicators of such sterling qualities are observations of actual performance in the role over an extended period of time.

Issue 10: Divergent Contextual or Situational Characteristics

During recent years, personnel from the Centre for Principal Development have participated in the delivery of in-service professional development programs for school administrators in a variety of locations across Canada. These sites are as diverse as Baker Lake and Iqaluit in the Northwest Territories as well as Toronto, Waterloo, Renfrew, Belleville and Simcoe in Ontario. In-service experiences in widely divergent locales has reinforced our belief that much of the knowledge and many of the generic skills required of school administrators are fairly universal whether the site is in a remote area of the Canadian Arctic or in the centre of a large cosmopolitan city. However, what is equally apparent is that often within a single community the contextual circumstances and operational conditions associated with a school can vary significantly. Sometimes the variation is in only one area, but with critical implications for practice. To state the point more clearly, the knowledge, skills and attitudes characteristic of effective school administrators are remarkably consistent across Canada and well supported by research findings. However, because of local circumstances of a contextual or situational nature, administrators may encounter more or less difficulty in developing the requisite skills, acquiring the necessary professional knowledge, or exercising school and system leadership.

The situational variations evident among schools and school districts compromises, at least potentially, the relevance of performance appraisal instruments which attempt to measure administrator effectiveness against fixed performance standards. Examples of our own experience help to illustrate the point.

Most Ontario school districts, particularly in rural areas where small schools are

common, still require some of their principals to teach full-time or some portion of the time. These teaching principals frequently report that they have difficulty reconciling the demands imposed on them by the Ministry and school district to be instructional leaders and policy implementors with the realities of their available administration time. Having their effectiveness as principals assessed by an instrument designed with full-time principals in mind simply adds insult to injury. On the other hand, there are many schools in regions of rapid population growth (e.g., Dufferin-Peel and York Region) where whole schools are housed in a series of portable classrooms or other temporary housing in chronically over-crowded or resource constrained conditions. These trying and very particular circumstances may not be compatible with the kinds of selection and appraisal instruments most designers have in mind.

Issue 11: Reactive and Non-Reflective Working Conditions:

The time constrained conditions under which most principals and superintendents work is probably the single most obvious incentive for the development of selection and performance appraisal instruments. The typical work practices of principals and superintendents are simply not compatible with the demands of most selection and appraisal procedures (see Leithwood & Montgomery, 1986 and Fullan, Park & Williams, 1988). The work day of most school administrators can probably be best characterized by rapid fire decisions, short encounters with a multitude of people with different interests and a generally broken field approach to getting things done. There is not a lot of time for planning or reflection much less time available for the systematic processing of instrument data in a formative appraisal situation.

It does not take much imagination to realize that interviewing candidates for an administrative position takes a lot of time; usually outside normal working hours. Visiting schools or classrooms to evaluate the performance of principals requires considerable time, particularly if the objective is formative goal setting as opposed to summative evaluation. There is, therefore, an obvious appeal for practitioners in an

instrument which promises to efficiently select job candidates or evaluate job performance with minimal or no personal interaction required.

Instrument developers state that instruments must be properly administered and should ideally be part of a comprehensive package which makes use of more traditional approaches to assessment. If one keeps the typical administrator's day firmly in mind, it is evident that simply recommending the use of their instrument along with other more traditional supporting procedures that require more time and interaction does nothing to minimize the risk of misuse of such instruments. Instrument developers have an ethical responsibility, which they may or may not want to recognize, to ensure that users do not simply use their instruments as "quick and dirty" solutions to their more fundamental time management problems. Unfortunately, once an instrument is in the hands of the user there is typically little the designer can do to ensure they are used appropriately.

Issue 12: User's Preferred Administrative Style

A considerable amount of research conducted during the last decade in Canada, the United States, Australia and some parts of Europe has addressed the notion of administrators' preferred work styles (see Leithwood and Montgomery, 1986; Fullan, Park & Williams, 1988; Hall, 1988; Rutherford, 1988). As previously discussed, Leithwood and Montgomery identify four administrative styles among principals; the Administrator, the Humanitarian, the Program Manager, and the Systematic Problem-Solver. Hall and Rutherford identify three distinct styles for change facilitators which correspond roughly with the first three styles of the Leithwood framework. In their recently published document on the supervisory officer in Ontario, Fullan, Park and Williams identify three dimensions to describe style variations among superintendents. These are system-driven versus school-driven, reflective versus firefighting, and generalist versus specialist.

The preferred work styles manifested by school administrators has implications for their potential use or misuse of selection and appraisal instruments. Let us consider

principal work styles and then superintendent work styles to illustrate these implications. As a result of using the Leithwood and Montgomery framework to profile large numbers of Ontario principals over a period of years, we can say with reasonable confidence that most practicing principals probably reflect the orientations and work styles best represented by the Humanitarian and Program Manager styles. As has been previously indicated, few principals can currently be described as the highly effective Systematic Problem-Solvers and a somewhat larger proportion as traditional Administrators.

Let us now consider how well these principals' preferred work styles relate to the use of performance appraisal instruments. The minority of principals with an Administrator style, focussed on traditional building management and "keeping up appearances", are unlikely to value the more formative aspects of evaluation. The more ubiquitous Humanitarian's preoccupation with maintaining good relationships among staff and students is probably not consistent with the summative focus of performance appraisal instruments. The potential for conflict and unpleasantness resulting from such instrument use is simply too high. Formative types of evaluation probably have more appeal to the Humanitarian because at least this type of evaluation may contribute to good staff relationships and job satisfaction through improved performance. Principals whose performance reflects the two highest levels of the Leithwood framework could be expected to be ready consumers of data from performance appraisal instruments. They are systematic in their approach to school administration, focussed on the needs of programs or students, and generally highly effective administrators who could be expected to consider their own performance appraisal process a priority. From this discussion it is apparent that principals with preferred work styles may or may not be willing or comfortable with the use of performance appraisal instruments.

In the case of superintendents, Fullan, Park and Williams (1988) found that the majority of superintendents are system oriented rather than school oriented. They are

also more inclined to a reactive or firefighting approach than to a reflective and proactive approach to administration. The former orientation suggests a managerial orientation likely to be supportive of the use of summative instruments for selection and appraisal because they enforce uniform standards for personnel and foster accountability. Instruments which are efficient in terms of the time and effort required for implementation are similarly appealing because they leave more time for crises management and lessen the need for sustained one on one contact with personnel which is not characteristic of the managerial style.

(d) Links to Factors Influencing Use

Listed above are several issues that face both instrument developers and educational practitioners alike concerning their involvement with instrumentation for principal selection and appraisal. These may be thought of as factors that bear some influence on the nature and extent of use that may be expected. In the absence of empirical data, we attempt to draw links between the issues listed above and the Cousins-Leithwood framework for evaluation utilization. Table 1 lists the research-based factors and our estimation of which of the issues described above is associated with each. Also listed, at the risk of speculation, are projected misutilization and non-utilization outcomes that seem likely should conditions permit.

A quick glance at Table 1 shows that many of the speculative, likely outcomes are classified as misevaluation associated with instrument developer issues. That is to say, if the instrument is lacking in demonstrated construct validity, predictive validity (especially if used for selection purposes), content validity, reliability, and so forth, the most likely result is that the score-based inferences will be rendered uninterpretable. This was also thought to be a likely outcome if normative data were not provided to the user because there would be no explicit basis upon which to establish performance standards. Misevaluation might also occur if scores are calculated incorrectly due to poor administrative instructions.

Table 1

Factors by Instrument Developer/Practitioner
Issues by Probable Misuses/Non-Use Outcomes

FACTOR	ISSUE[a]	PROBABLE MISUSE OR NON-USE OUTCOME
EVALUATION IMPLEMENTATION FACTORS		
1. Evaluation Quality	5. Reliability 6. Error of Estimation	Misevaluation
2. Credibility	1. Construct Validity 2. Predictive Validity	Misevaluation
3. Relevance	3. Content Validity	Misevaluation
4. Communication Quality	4. Normative Data	Misevaluation
5. Findings	7. Responsibility for Use	Misuse, Abuse
6. Timeliness	7. Responsibility for Use	Unintentional Non-Use Misevaluation
DECISION OR POLICY SETTING CONTEXT FACTORS		
1. Information Needs	8. Evolving Role of School Administrators 9. Using Predetermined Selection/Appraisal Criteria 10. Divergent Contextual or Situational Characteristics 11. Reactive and Non-Reflective Work Conditions	Misevaluation Justified Non-Use
2. Decision Characteristics	11. Reactive and Non-Reflective Work Conditions	Misevaluation Unintentional Non-Use
3. Political Climate	11. Reactive and Non-Reflective Work Conditions	Abuse, Misuse
4. Competing Information	10. Divergent Contextual or Situational Characteristics	Justified Non-Use
5. User Personal Characteristics	12. User Preferred Administrator Style	Abuse
6. User Commitment and/or Receptiveness	11. Reactive and Non-Reflective Work Conditions 12. User Preferred Administrator Style	Abuse

[a] Instrument developer issues (1-7); Practitioner issues (8-12).

All of the above considerations correspond to the responsibility of the instrument developer but the user can also be responsible for misevaluation. Consider the case where the technical quality of the information is adequate but the data are not made available in time to support decisions. This problem can hardly be attributed to the instrument developer. Looked at from the perspective of the decision-maker (e.g., selection committee member) who requires the data, the outcome might be construed as unintentional non-use according to the Alkin & Coyle (1988) classification system. Two additional undesirable outcomes associated with evaluation (instrument) implementation factors are much less likely but attributable to the nature of the findings. Should instrument scores show unexpected or undesirable outcomes there may be a tendency for users to (1) surpress their revelation (abuse) or (2) deliberately manipulate them (misuse) for a purpose such as personal gain. A selection situation where an undesirable candidate scored high is an example of this kind of circumstance.

Instrument scores that derive from restrictive criteria and that are not sensitive to regional or local differences, nor the changing role of principals do not meet the needs of the educational practitioner and will likely lead to one of two possible outcomes. First, if used for decision-making or performance improvement purposes misevaluation is likely. Second, if intentionally not used due to the deficiencies outlined above, the result is justified non-use according to the Alkin-Coyle framework. A decision by a selection committee member to reject instrument scores known to be of low validity is certainly justifiable. These outcomes assume, of course, that the instrument has been rendered significantly obsolete and that no other complimentary data that make up for the deficiencies are available.

The issue concerning the reactive nature of the milieu in which educators work can lead to a variety of undesirable outcomes associated with several decision or policy setting factors. Reactive and non-reflective work conditions imply that decisions are made in haste and that data, if processed, may be processed incorrectly by users (misevaluation). Alternatively, they may not be processed due to lack of time for

reflection and unintentional non-use is the result. Further, attitudes towards such evidence as that provided by instruments may be sufficiently poor as to result in the intentional non-use (abuse) of data by users.

Divergent contextual and situational characteristics might render an instrument inappropriate, as mentioned, and as such it would likely not be politically prudent to make use of it resulting in either efforts to surpress or modify the data. Perhaps a more likely outcome of this contextual situation would be reliance on other competing sources of information that are viewed as more appropriate in which case non-use of instrument data would be justified. For example, an appraiser might choose to rely more on personal observations of an appraiser's peers, staff, community, and the like.

Finally, preferred administrator styles that are not conducive to systematic data collection and use correspond to user personal characteristics and attitudes that would mitigate against intended uses of instrument data. The most likely result of these styles is intentional non-use or, in Alkin and Coyle's terms, abuse. An appraisee might choose to ignore indications of areas for growth highlighted by instrument data, for example.

The above analysis cannot be considered conclusive in any sense of the word. Its lack of grounding in empirical data and highly speculative nature most certainly preclude this outcome. What the analysis accomplishes, however, is that it uncovers a new way of thinking about possible undesirable (and sometimes desirable as in the case of justified non-use) outcomes when instruments are used for selecting and appraising principals. It should be noted, however, that conspicuously absent from Table 1 is the term ideal use. This outcome, it is assumed, will occur if the condition of all factors is favourable. In the absence of direct data, it is difficult to estimate what combination of favourable and unfavourable conditions is required before ideal use can be expected. From an indirect perspective, however, Cousins and Leithwood (1986; Cousins forthcoming) found that factors associated with the decision or policy setting were most influential concerning utilization as education (conceptual development) whereas a mix

of factors from both dimensions of the framework (i.e., evaluation implementation and decision or policy setting) were shown to influence use for instrumental purposes. Perhaps one could extrapolate and suggest that personal commitment and flexibility on behalf of users would help to foster desired formative evaluation outcomes associated with instruments designed for selection and appraisal. The merit that this extrapolation deserves, of course, remains an empirical question.

Safeguards

Given the above conjecture regarding likely misuses of instrument data we are now in a position to recommend various safeguards that may be taken by instrument developers and practitioners alike. An important assumption we make is that the user or practitioner has the ultimate responsibility for use but that the instrument developer is ethically responsible for providing as much guidance as possible to ensure that appropriate uses are made.

Given this assumption an obvious first step would for instrument developers be to produce a technical manual that:

- explicitly states the purposes and applications for which the the instrument is recommended (especially if the instrument is designed specifically for formative or summative purposes);
- describes clearly theoretical underpinnings (e.g., conception of effectiveness in the role of principal) on which the instrument development was based;
- indicates that instrument scores describe but do not explain observed levels of performance;
- stresses that the use of the instrument alone will not provide all of the relevant facts of a description, and that the collection of other complimentary data is strongly recommended;
- clearly explains the how to properly score the instrument;
- provides evidence about reliability and various types of validity, in addition to normative information.

- indicates the "expiry date" of the instrument or when its content is expected to be updated according to changes in the conceptions of effective practice.

Some of these suggestions are due to Brown (1980) in his discourse on guidelines for test use. As a minimum, such a manual would provide users with the relevant facts that they need to consider prior to using the an instrument for practical purposes.

Another step would be to develop a standard (e.g., computerized) system for processing the data and either make software available to users or offer a centralized processing service. This step would help eliminate errors in scoring attributable to misunderstood scoring instructions and simple hand calculation errors. Such errors are more likely for an instrument with multiple scales that will likely employ in scoring, weighted averages over the dimensions as opposed to simple summated ratings.

Instrument developers might also engage in extensive field training of practitioners in the proper administration and use of the instrument. Such field training might involve pilot testing or some form of mock administration so that subtle administrative features and quirks are more clearly understood by users.

Instruments might best be used in the context of administrative decision-making as part of a battery of instruments and procedures suited to selection and promotion to positions of added responsibility. Procedures that employ multiple raters, including the principal/vice principal or principal candidate are recommended. It is crucial that raters have a good knowledge of these individuals' working behaviour and performance characteristics. There also exists a great potential for instruments to be used for diagnostic purposes. For example, the comparison of data from multiple raters including various colleagues and a self assessment by the appraisee might serve as an excellent stimulant for discourse about current performance, obstacles to growth, strategies to overcome obstacles, and the like.

As a safeguard against the potential misuse of selection/appraisal instruments which may not reflect the realities of ever evolving administrative roles, users of such instruments, should be wary of claims that any instrument used in isolation can identify

the ideal practitioners required to meet the needs of education in the future. This is not to argue against rational selection processes at one level or another. The point is we cannot really be sure what educational needs will develop even five years from now. Furthermore, potential users of such instruments must ensure that they are based on the most current image of the role which is relevant to the particular needs of a school district. In short, instrument users must be discriminating consumers. They should critically assess the validity and currentness of the image for the role implicitly espoused by such an instrument and recognize that instruments will at best identify effective practitioners for current circumstances.

Significance/Conclusions

At a time in the history of Canadian education when school systems are confronted with unprecedented opportunities to affect change through the selection and appraisal of school administrators, the importance of choosing worthwhile instruments and effectively integrating them into system standard operating procedures is underscored. Failure of educators to give careful consideration to the crucial attributes of such instruments, the characteristics of the practical settings in which they will be implemented and the effects of such factors on the legitimate use of instrument data may result in inappropriate uses, misuses or abuses: this has severe implications not only for individual careers but for systems' desires to improve the effectiveness of their administrators.

References

Alkin, M.C. & Coyle, K. (April, 1988). *Thoughts on evaluation misutilization (and associated topics).* Paper presented at the annual meeting of the American Educational Research Association, New Orleans

Alkin, M.C., Daillak, R. & White, B. (1979). *Using evaluations.* Beverly Hills, CA: Sage Publications.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (1985). *Standards for educational and psychological testing.* Washington, DC: APA.

Bernardin, H.J. & Beatty, R.W. (1984). *Performance appraisal: Assessing human performance at work.* Boston, MA: Kent.

Brown, F.G. (1980). *Guidelines for test use: A commentary of the standards for educational and psychological tests.* Washington, D.C.: National Council on Measurement and Evaluation (NCME).

Cronbach, L.J. (1980). Validity on parole: How can we go straight? *New Directions in Testing and Measurement, 5,* 99-107.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education

Cousins, J.B. (forthcoming). *Factors influencing principals' use of performance appraisal data.* Doctoral dissertation. Toronto: University of Toronto, Ontario Institute for Studies in Education.

Cousins, J.B. & Leithwood, K.A. (November, 1987). Development of a Principal Profile-based Instrument (PPBI) for the selection and appraisal of school administrators. Paper presented at the conference, *The Selection of School Administrators*, Toronto.

Cousins, J.B. & Leithwood, K.A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56* (3), 331-364.

Duke, D.L. & Stiggins, R.J. (1985). Evaluating the performance of principals: A descriptive study. *Educational Administration Quarterly, 21* (4), 71-98.

Fullan, M.G. (1988). *What's worth fighting for in the principalship?* Toronto: Ontario Public School Teachers' Federation.

Fullan, M.G., Park, P.B. & Williams, T.R. (1987). *The supervisory officer: Current practice and recommendations for the future.* Toronto: The Queen's Printer for Ontario.

Hall, G. (April, 1988). Development of a framework and measure (in English) for assessing principal change facilitator style. Paper presented as the annual meeting of the American Educational Research Association, New Orleans.

Hallinger, P. & Murphy, J.F. (1987). Assessing and developing principal instructional leadership. *Educational Leadership. 45* (1), 54-61.

Hallinger, P. & Murphy, J.F. (1984). Assessing the instructional management behaviour of principals. *The Elementary School Journal. 86* (2), 217-247.

Huberman, M. (1987). Steps toward an integrated model of research utilization. *Knowledge: Creation, Dissemination, Utilization, 8* (4), 586-611.

Kennedy, M.M. (1983). Working knowledge. *Knowledge: Creation, Dissemination, Utilization, 5* (2), 193-211.

Kennedy, M.M. (1984). How evidence alters understanding and decisions. *Educational Evaluation and Policy Analysis, 6* (3), 207-226.

Landy, F.J. & Farr, J.L. (1983). *The measurement of work performance.* New York, NY: Academic Press.

Lawton, S., Hickox, E., Leithwood, K. & Musella, D. (1986). *The development and use of performance appraisal of certified education staff in Ontario school boards.* Toronto, Ont.: The Queen's Printer for Ontario.

Leithwood, K.A. & Begley, P. (1985). *School management in Canada: A description and analysis of selected issues.* Technical report prepared for the International School Improvement Project (OECD - CERI).

Leithwood, K.A. & Montgomery, D. (1986). *Improving principal effectiveness: The Principal Profile.* Toronto: OISE Press.

Leithwood, K.A. & Montgomery, D. (1982). The role of the elementary principal in program improvement. *Review of Educational Research, 52,* 309-339.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10* (9), 9-20.

Musella, D & Lawton, S. (1986). *Selection and promotion procedures in Ontario school boards.* Toronto: The Queen's Printer for Ontario.

Patton, M.Q., Grimes, P.S., Guthrie, K.M., Brennan, N.J., French, B.D., & Blyth, D.A. (1977). In search of impact: An analysis of the utilization of federal health evaluation research. In C.H. Weiss (Ed.)., *Using social research in public policy making.* (pp. 141-163). New York: D.C. Heath.

Patton, M.Q. (1988). In M.C. Alkin (Ed.)., *Evaluation reflections.* CSE Technical Report. Los Angeles: UCLA, Center for the Study of Evaluation (in press). {Cited by Alkin & Coyle (1988)}.

Peterson, K. D., Marshall, C. & Grier, T. (1987). Academies for assistant principals. *Educational Leadership, 45* (1), 47-48.

Pitner, N.J. & Hocevar, D. (1987). An empirical comparison of two-factor versus multifactor theories of principal leadership: Implications for the evaluation of school principals. *Journal of Personnel Evaluation, 1* (2), 93-109.

Robinson, G. E. & Block, A.W. (1982). The principal and achievement: A summary of 22 studies. *Principal, 62* (2), 53-57.

Rutherford, W.L. (April, 1988). Distinguishing three change facilitator styles from analysis of principal interventions. Paper presented as the annual meeting of the American Educational Research Association, New Orleans.

Schmitt, N., Noe, R., Meritt, R., Fitzgerald, M., & Jorgensen, C. (1983). *Criterion-related and content validity of the NASSP assessment center.* Reston, VA: NASSP

Stiggins, R.J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice,* Fall, 33-42.

Trider, D.W.A. & Leithwood, K.A. (1988). Influences on principals' practices. *Curriculum Inquiry, 18* (3), 289-311.

Vandenberghe, R. (April, 1988). Development of a framework and measure (in Dutch) for assessing principal change facilitator style. Paper presented as the annual meeting of the American Educational Research Association, New Orleans.

Weiss, C.H. (1981). Measuring the use of evaluation. In J.A. Ciarlo (Ed.)., *Utilizing evaluation: Concepts and measurement techniques* (pp. 17-33). Beverly Hills, CA: Sage.

Yalow, E.S. & Popham, W.J. (1983). Content validity at the crossroads. *Educational Researcher, 12* (8), 10-14.